Accepted Manuscript

Web based health surveys: Using a Two Step Heckman model to examine their potential for population health analysis

Karyn Morrissey, Peter Kinderman, Eleanor Pontin, Sara Tai, Mathias Schwannauer

PII: S0277-9536(16)30343-4

DOI: 10.1016/j.socscimed.2016.06.053

Reference: SSM 10726

To appear in: *Social Science & Medicine*

Received Date: 14 November 2015

Revised Date: 18 May 2016

Accepted Date: 29 June 2016

Please cite this article as: Morrissey, K., Kinderman, P., Pontin, E., Tai, S., Schwannauer, M., Web based health surveys: Using a Two Step Heckman model to examine their potential for population health analysis, *Social Science & Medicine* (2016), doi: 10.1016/j.socscimed.2016.06.053.

Manuscript #: SSM-D-15-03161

Manuscript Title: "Web based health surveys: Using a Two Step Heckman Model to examine their potential for population health analysis"

Authors:

Karyn Morrissey, University of Exeter Truro, UK

Peter Kinderman, University of Liverpool, UK

Eleanor Pontin, University of Liverpool, UK

Sara Tai, University of Manchester, UK

Mathias Schwannauer, University of Edinburgh, UK

Corresponding Author:

Karyn Morrissey,

Senior Lecturer

University of Exeter

European Centre for Environment and Human Health, Knowledge Spa

Truro, TR13HD, UNITED KINGDOM

k.morrissey@exeter.ac.uk

Web based health surveys: Using a Two Step Heckman Model to examine their potential for population health analysis

Abstract

In June 2011 the BBC Lab UK carried out a web-based survey on the causes of mental distress. The 'Stress Test' was launched on 'All in the Mind' a BBC Radio 4 programme and the test's URL was publicised on radio and TV broadcasts, and made available via BBC web pages and social media. Given the large amount of data created, over 32,800 participants, with corresponding diagnosis, demographic and socioeconomic characteristics; the dataset are potentially an important source of data for population based research on depression and anxiety. However, as respondents self-selected to participate in the online survey, the survey may comprise a non-random sample. It may be only individuals that listen to BBC Radio 4 and/or use their website that participated in the survey. In this instance using the Stress Test data for wider population based research may create sample selection bias. Focusing on the depression component of the Stress Test, this paper presents an easy-to-use method, the Two Step Probit Selection Model, to detect and statistically correct selection bias in the Stress Test. Using a Two Step Probit Selection Model; this paper did not find a statistically significant selection on unobserved factors for participants of the Stress Test. That is, survey participants who accessed and completed an online survey are not systematically different from non-participants on the variables of substantive interest.

Keywords: United Kingdom; Depression; web-based surveys; sample selection bias; Two Step Probit Selection Model

1. Introduction

The fast pace of technology changes and the increasing ease of developing online surveys via companies such as survey monkey has resulted in web surveys attracting considerable interest from academic researchers (Fleming and Bowden, 2009; Strabac & Aalberg, 2011). Internet surveys are cost efficient and provide a wide range of new possibilities for data collection and for incorporation of complex visual information into a questionnaire (Strabac & Aalberg, 2011; Fleming and Bowden, 2009). Table 1 presents an overview of the advantages and disadvantages associated with Internet surveys. The main drawback commonly cited with web-based surveys is potentially low levels of access to, and use of Internet services across the general population (Bethlehem, 2010). However, Internet coverage and access is improving rapidly across all demographics (Bethlehem, 2010; Strabac & Aalberg, 2011; ONS, 2015). In England, a recent survey on internet usage by the Office of National Statistics (ONS, 2015) found that between January and March 2015, 86.2% of the English population had used the Internet in the last 3 months (recent users), while 2% (1.1 million) had last used the Internet more than 3 months ago and 11% (5.9 million) had never used it. Regionally, it was found that the South East of England reported the highest level of Internet usage (89.6%) and the North East the lowest (81.6%).

Table 1 Advantages and disadvantages of online surveys

However, even if an individual has Internet access, not all people are equally reached by Internet solicitation. The majority of web surveys allow participants to self-select into the sample (Schonlau et al, 2004; Bethlehem, 2010). Respondents are those people who happen to visit the website and then decide to participate in the survey (Bethlehem, 2010). The decision to participate may depend on personal characteristics correlated with the subject

matter or how the survey is hosted and subsequently advertised. This is in contrast to the theory of probability sampling where each individual must have an equal chance of being included in the survey for the sample to generate unbiased results (Kish, 1965; Bethlehem, 2010; Sanchez-Fernandez et al., 2012). Self-selection into a survey implies that the principles of probability sampling are not followed (Berk, 1983; Lee and Marsh, 2000; Bethlehem, 2010). When this situation arises it is referred to as sample selection bias (Rubin, 1976; Berk, 1983). Selection bias occurs because non-participation is rarely random (e.g., distributed equally across subgroups); instead, bias often is correlated with variables that also are related to the dependent variable of interest (Goodfellow et al., 1988; Winship and Mare, 1992). Sample selection bias is a threat to both the internal and external validity of the model (Berk, 1983; Lee and Marsh, 2000; Cuddleback et al., 2004) whereby independent variables are correlated with a disturbance term (i.e., error) and analyses based on biased samples can lead to inaccurate estimates of the relationships between variables (e.g., regression coefficients).

In theory, non-probability samples may yield results that are just as accurate as probability samples. This situation may arise if the factors that determine a population member's presence or absence in the sample are all uncorrelated with the variables of interest in a study (Lee and Marsh, 2000; Yeager et al., 2011). The possibility of representative data from online surveys needs to be formally tested before such data may be used for research purposes. A method is therefore required that allows examination of whether the online survey of interest suffers from selection bias or unobserved heterogeneity before the data may be used for population based research. The conventional approach to missing data in a survey, the imputation approach has two important limitations if sample selection bias is present in the survey (Hogan et al., 2012). First, imputation assumes that no unobserved variables associated with dependent variable influence participation in the survey: non-

participants are missing at random. However, as noted above this may not be the case. Second, it ignores regression parameter uncertainty in the imputation model, resulting in confidence intervals (CI) that are too small (Hogan et al., 2012). The use of data imputation in the presence of sample selection bias is equally as invalid as using the raw data without imputation.

Beginning in the late 1970s (Greene, 1981; Heckman, 1976, 1978, 1979), methods for detecting and statistically correcting selection bias were developed in economics and related areas. In the decades since, as noted by Cuddleback and colleagues (2004), an extensive literature has evolved in the area of sample selection bias (Berk, 1983; Lee & Marsh, 2000; Stolzenberg & Relles, 1997). These methods are known as sample selection models and although developed over forty years ago, they are designed to overcome the same issues associated with Internet based surveys in the New Millennium. Using a variant of Heckman's (1976) two-step selection model (Van de Ven and Van Pragg, 1981), this paper examines whether data collected on mental distress at a UK University based Institute of Psychology and the BBC Lab UK via a web-based survey suffers from sample selection bias. If sample selection bias is not found, the Stress Test may be used to examine the prevalence of depression in the general population and other socio-economic analyses. However, if sample selection bias is found the simultaneous estimation of both models allows for the correction of the estimated coefficients in the outcome model. While the outcome of this research is of interest to the researchers involved in the original research project (Kinderman et al., 2013), such an analysis also extends the literature on the appropriateness of using online survey data for public health research.

Section 2 introduces the BBC Lab UK, the Stress Test and some descriptive statistics to examine the representativeness of the survey relative to external data sources. Section 3

introduces a Heckman style model for sample selection bias, used for binary variable analysis, the Two Step Bivariate Probit Model. Section 4 presents the results of a Two Step Bivariate Probit Model (Van de Ven and Van Pragg, 1981) used to test the Stress Test for sample selection bias. Section 5 offers a discussion of the results within the context of the usefulness of online surveys in public research.

2. The BBC, BBC Lab UK and the Stress Test

The British Broadcasting Corporation (BBC) is the public service broadcaster of the UK and the world's oldest national broadcasting organization. Originally providing radio services, the BBC service now includes television, radio and online platforms, which are available both nationally and internationally (Kinderman et al., 2015). In response to other online 'citizen science' projects the BBC Lab UK was launched in 2009. The BBC Lab UK invited members of the public to take part in science experiments investigating different aspects of psychology, sociology and health by completing tests and surveys online. The website was active form 2009 to 2013 until data collection ceased in May 2013 and the results for each experiment remain online (http://www.bbc.co.uk/labuk). Academic researchers designed each web experiment (see for example Savage et al., 2013; Kinderman et al., 2015; Rentfrow et al., 2015). Each experiment was structured to give feedback to the participant, immediately after they had submitted their data. As part of the BBC Lab UK and in collaboration with an Institute of Psychology in the UK, the 'Stress Test' was launched on 'All in the Mind' a BBC Radio 4 programme in June 2011 (Pontin, 2012; Kinderman et al., 2011). The test's URL was publicised on BBC Radio 4 and made available via BBC web pages and social media. Visitors to the Stress Test's homepage accessed the test by signing in using BBC online membership (Kinderman et al., 2015).

All persons gave their informed consent prior to their inclusion in the study (Kinderman et

al., 2011). On accessing the test webpage, participants were presented with the option of starting the test or following links to gain further information. The test was 'live' for a 4-week period (Kinderman et al., 2015). Participants were required to complete the test in one sitting and the test took approximately 20 minutes to complete (Pontin, 2012). Participants could opt out of the survey at any point. To minimise variation in survey responses, items were completed in a fixed order from a dropdown box and some tasks had time limits (Kinderman et al., 2015). Once the test was completed, participants were not permitted to complete the test again (Pontin, 2012). Only individuals over the age of 18 years could complete the Stress test. Over 32,000 individuals responded to the online Stress Test, of which over 5,000 were from outside the UK.

The first part of the Stress Test gave an overall measure of people's current mood using the Goldberg Anxiety and Depression Scale (GADs) (Goldberg et al., 1988) and the BBC Well-being Scale, a new measure of general well-being developed for The Stress Test (Kinderman, et al., 2011). Developed from a 36 items of the Psychiatric Assessment Schedule (Surtees et al., 1983), GADs is eighteen-item self-report symptom inventory with 'yes' or 'no' responses to items asking how respondents had been feeling in the past few weeks. Nine items each comprise the anxiety and depression scales and the scales can be separated to examine the case prevalence of depression (GDs) or Anxiety (GAs) alone. Researchers have concluded the GADs to be a valid and acceptable method of detecting depression and anxiety (Smith, 2004; Montero-Marin et al., 2014; Koloski et al., 2008). Recent international research using the GADS includes research on menopause in Spain (García-García et al., 2015), expectant fathers in Australia (Leach et al., 2015), obesity in Latin America (Blümel et al., 2015) and the indoor environment in Italy (Magnavita, 2015).

Depression and anxiety frequently coexist in the same individual (Christensen et al., 1999),

either concurrently or at different times, and numerous studies show that the presence of an anxiety disorder is the single strongest risk factor for development of depression (Hranov, 2007). However, controversy continues over the nature of the relationship between depression and anxiety, some believing they are distinct, separate entities while others (the majority) view them as overlapping syndromes that present at different points on a phenomenological and/or chronological continuum (Hranov, 2007). However, the aim of this paper is to provide a pedagogical illustration of the Heckman sample selection model within the context of the new data landscape available to researchers. Thus, the relationship between the dependent and independent variables is of lesser interest to the authors. Importantly, the nine item anxiety and depression components of the GADs, the GDs and GAs have been used as separate scales within research to date. For example, Sforza et al., (2016) used the Gas component of the GADS to examine anxiety and sleep apnea.

Thus, for simplicity this paper therefore focuses on examining the depression component of the GADS. Individuals receive a score along a nine-point scale, where a score of 1 is a low probability of being diagnosed with case depression and 9 is a high probability of being diagnosed with case depression. Including the GADs and the BBC Well-being Scale, the Stress Test consisted of 12 sections and included; demographic and socio-economic questions; familial mental health; social inclusion; negative life events (historic and recent); neurocognitive responses to negative feedback and positive and negative stimuli; and psychological processes (response style and attribution style). Participants were also asked to write, using free text, their first and second biggest cause of stress (Pontin, 2012). Table 2 provides descriptive statistics for respondents to the Stress Test who reside in England. Descriptive statistics presented by Kinderman et al., (2011) found that Stress Test respondents were predominantly white, had slightly higher earnings, and were better educated than the general population.

Table 2 Descriptive Statistics for English Respondents to the Stress Test

Geographical Data

The Stress Test also asked participants for the first part of their postcode (referred to as the outward code). For example, if a respondent lived in L1 9AD (a Liverpool address) they supplied the L1 component of their postcode. Over 90% of respondents filled out their postcode and those that did not were removed from the Stress Test sample. Data processing in the statistical software Stata (StataCorp, 2013) linked the first part of each respondent's postcode to the National Statistics Postcode Lookup (NSPL) table produced by ONS Geography. This process allowed a range of higher geographies including ward, Lower Super Output Area (LSOA) and Government Office Region and urban/rural classification (ONS, 2014), to be merged with the Stress Test data. As postcodes were merged to higher-level geographies, particularly the LSOA the exclusion of the second component (referred to as the outward code) becomes irrelevant.

The inclusion of the LSOA variable also provided the spatial referencing required to link data from the index of multiple deprivation for England (IMD) 2010 (Department for Communities and Local Government, 2010) to the data from the Stress Test. A socio-spatial index, the IMD uses a combination of Census data and data derived from other sources such as the Inland Revenue, the Department of Health and the Department of Transport to measure the multiple facets of deprivation at the small area level (Morrissey et al., 2016). Linking the IMD for England 2010 to data from the Stress Test was important for two reasons, one conceptual, one methodological. Conceptually, place of residence is strongly patterned by social position and there is a growing appreciation of the affect neighbourhood characteristics can have on individual health outcomes (Diez Roux & Mair 2010). The IMD

2010 provides a way in which the impact of area level deprivation can be accounted for in health research (Barnett et al., 2012; Morrissey et al., 2016).

The IMD is available in two numerical forms; as a rank variable, which shows how an individual LSOA compares to other LSOAs in the country, and as an absolute score (Morrissey et al., 2016). The IMD 2010 was constructed by combining seven general welfare domain scores weighted as followed; income (22.5%), employment (22.5%), housing and disability (13.5%), education, skills and training (13.5%), barriers to housing and services (9.3%), crime (9.3%), and living environment (9.3%). The IMD is based on data at the Lower Super Output Areas (LSOA) with each LSOA containing on average 1500 people or 600 households (Morrissey et al., 2016). The LSOA ranked number one is the most deprived, with higher rankings indicating less deprived areas. Overall, the Stress Test contains a comprehensive collection of demographic, socioeconomic, clinical, psychological and spatial data that if found to be robust in terms of sampling is a powerful dataset for further public health research. However, due to the manner in which the survey was (a) launched, via a BBC 4 Radio programme and (b) hosted, only on the BBC website, individuals with a higher propensity to listen to BBC Radio 4 may therefore exhibit a higher participation rate compared to the general population. A report commissioned by BBC Radio in 2010 (BBC Radio Review, 2010) found that average listenership to BBC Radio 4 was disproportionately higher in the Greater London region (24.6%) and the South East region (27%) compared to national listenership rates of 19%. The advertisement of the Stress Test on BBC Radio 4 and the above average listenership to BBC Radio 4 in the southeast region may mean that there are a disproportionality higher number of respondents from this two regions, thus creating a regionally biased sample.

With regard to spatial representativeness of the Stress Test, it was found that 43% (14,124)

of LSOAs in England did not have a single survey respondent. However, this is not necessarily an issue of concern. The collection of a geographically representative sample does not require that each geographical unit have a respondent. Instead, surveys are designed to yield a representative sample of the population by setting demographic, socio-economic and spatial quotas based on national registers (usually the Census of Population). Initial Chi-Squared analysis found no statistical regional bias (90% significance level) between the respondents to the Stress Test compared to the Health Survey of England. However, sample representativeness at the sub regional level, particularly once demographic and socio-economic factors are accounted for may still be an issue for the Stress Test.

3. Modelling Sample Selection

Heckman style models are based on the simultaneous estimation of two multiple regression models, an outcome equation and a selection equation (Bushway et al., 2007; Barnighausen et al., 2011). The simultaneous estimation of the outcome and selection equations allows for any correlation between the unobserved error terms for the dependent variable and participation in the survey (Barnighausen et al., 2011). In choosing whether it is appropriate to use Heckman type models to investigate sample selection bias, the data under analysis must met a number of criteria. These criteria include:

- A full set of observations for each variable for both participants and non-participants
- A dependent variable in the selection model that is an appropriate proxy for survey participation and non-participation
- The selection of an appropriate exclusion variable in the selection model

If the dataset meets these criteria only then can a Heckman style model be implemented. Indeed, it is important to note that the Heckman model is a very sensitive model and the authors found that if the above criteria are not met the model simply will not run or converge. The next three sections explore these criteria, with reference to the data requirements for the successful implementation of a Heckman Style Sample Selection model and data from the Stress Test.

Observations on participants and non-participants

With the exception of the dependent variable in the outcome model, selection models can only be used when there is a full set of observations for participants and non-participants for each variable (Geneletti et al., 2011; Barnighausen et al., 2011). Linking the Stress Test with external geo-referenced data that includes the IMD score, region and urban/rural classification for all LSOAs in England provides data for both participants and non-participants of the Stress Test. This paper therefore follows Chaix et al., (2011) who sought to identify selective participation in a cohort study using data on the neighbourhood in which participants or non-participants resided. Creating a dataset with area level data for both participants and non-participants and using a Heckman style Sample Selection model allows us to examine whether there is a difference between respondents to the web-based survey who accessed and completed the online survey in full and individuals that did not access the survey.

Figure 1 Heckman Style Sample Selection model and data from the Stress Test

Choosing a proxy for survey participation: the Selection Equation

As the decision being modelled is an individual's choice to participate in a survey or not, the dependent variable for the selection equations must be a binary (Barnighausen et al., 2011).

A report commissioned by BBC Radio in 2010 (BBC Radio Review, 2010) found that average listenership to BBC Radio 4 was disproportionately higher in the Greater London region and the South East region. The higher than average listenership to BBC Radio 4 in the southeast region of England suggests that residing in the southeast of England may be used as proxy for BBC Radio 4 listenership in the selection equation.

Exclusion Restriction

Even when data for participants and non-participants and an appropriate dependent variable for the selection model are available, research evidence demonstrates that the Heckman approach can seriously inflate standard errors if there is collinearity between the correction term and the included regressors (Moffitt 1999; Stolzenberg and Relles 1990). To ensure non-collinearity between the outcome equation and the selection equation, the selection equation must include an observed variable, z_i that affects why individuals may select to participate in study but does not influence the outcome variable (Bushway et al., 2007; Sartori, 2003). This variable is referred to as the exclusion restriction (Sartori, 2003). The identification of a valid exclusion restriction involves three steps (Bushway et al., 2007; Barnighausen et al., 2011). First, one must consider which of the variables available in a survey could be associated with survey participation. Second, it is important that the exclusion variables, z_i are not relevant predictors of the dependent variable in the outcome equation (depression). This second criteria eliminates an overwhelming number of variables (Bushway, 2007). Third, one must test whether the selection variable is indeed significantly associated with survey participation in a selection model, controlling for other observed variables.

Given the higher than average rates of BBC 4 listenership in the Southeast of England (BBC Radio Review, 2010), the newly created Southeast of England variable is used as a proxy for

listening to BBC Radio 4 and becomes the survey participation variable. A large number of satellite towns have developed in the South East region to accommodate high rates of commuting into London and the region has become highly urbanised. Within this context, this paper proposes using a rural residency as the exclusion criteria, z_i, for the selection equation. Following the criteria set out by Bushway (2007) on identifying a valid exclusion restriction, rural residency was chosen, as we believe that rurality will be:

a. Significantly, negatively associated with the South East region, the survey participation variable and;
b. Unrelated to the outcome variable, depression.

The 2011 ONS rural-urban classification (RUC2011) was used to designate if an LSOA was urban or rural. Using the predefined 2011 RUC2011 definition of rurality allows for a consistent definition of rural/urban LSOAs across datasets and as such is used as the de facto measure of rurality in the majority of health research in the UK (Kyte and Wells, 2010). The hypothesised insignificant relationship between rural residency and depression is based on research in England that found that once demographic and area level socio-economic factors (such as the Index of Multiple Deprivation) are accounted for, rural residency does not have a significant impact on outcomes of depression (Paykel et al., 2000; McKenzie et al., 2013). Following the recommended steps in specifying the selection model (Bushway et al., 2007) a probit model was used to test if there was a significant relationship between depression and residing in a rural area. It was found that controlling for IMD score, residing in a rural area is not associated with depression for respondents to the Stress Test. The urban/rural variable becomes the exclusion criteria in the selection model, the depression variable becomes the dependent variable in the outcome equation, while the IMD score of each of 32,000 LSOAs becomes an explanatory variable in both the outcome and selection models.

It should also be noted that a number of variables were tested as potential exclusion criteria. The initial choices included occupation type, as the Southeast region has a higher number of residents in professional occupations. However, professional occupation was significantly (positively) associated with outcomes of depression. Given the inconclusive association between ethnicity and depression (Mallinson and Popay, 2007), the ethnicity variable was classified as a binary variable, white or other, and also tested as a potential exclusion criterion. However, within this sample white ethnicity was significantly (positively) associated with outcomes of depression.

Specifying the Outcome and Selection Equations for the Stress Test

In most respects the outcome model is no different from any other multiple regression model and continuous, binary, multi-categorical, or other types of dependent variables may be used as dependent variables (Bushway et al., 2007). The majority of research using clinical scales to explore levels of depression and anxiety in the general population simply assumes that a diagnosis or scale can be applied with equal validity across a wide age range (Jorge et al., 2005). However, Jorm et al., (2005) found this issue becomes most serious in the very elderly where there is a high prevalence of physical disorders which can produce changes in some depression and anxiety symptoms. Research on depression and anxiety using the GADS (Jorm et al., 2005; Christensen et al., 1999) found that depression and anxiety symptoms decline across age groups. Given the importance of age as an exploratory variable for depression and the use of the GD scale in the Stress Test, this research follows Jorm et al., (2005) and reassigns the outcome variable, depression, as a binary variable according to the age and gender cut-offs presented in Table 3. For example, male respondents to the Stress Test aged 20-24 with a score of six or greater on the GDS were designated as case prevalent for depression and assigned a value of 1. Similarly female

respondents to the Stress Test aged 25-44 were assigned a value of 1 if they scored 5 or higher on the GDS.

Table 3 Cut-offs for Goldberg Depression by Age and Gender

The reassignment of the GDS as a binary variable according to age and gender specific cut-offs means that the traditional Heckman Two Stage Sample Selection Model for continuous variables cannot be used as part of this research. As such, this paper uses a Two Step Bivariate probit model to measure potential sample selection bias. This newly created depression variable, where 0 represents individuals not case prevalent and 1 represents individual's case diagnosed with depression, becomes the dependent variable in the outcome equation. The Two Step Bivariate Probit Model begins with the specification of the outcome equation, a probit model specified as follows:

$$y_i^* = \beta x_i + \varepsilon_i$$

$$y_i = 1 \; if \; y_i^* > 0, y_i = 0, otherwise$$

where y_i^* is an unobserved latent variable that determines the likelihood of a respondent to the Stress Test being diagnosed with depression *i*, and y_i^* depends on a vector of observed characteristics x_i and random error ε_i. Actual diagnosis of depression y_i is either negative (0) or positive (1), depending on whether y_i^* is above or below zero.

The dependent variable for the selection equations must be a binary variable, as the decision being modelled is an individual's choice to participate in a survey or not (Barnighausen et al., 2011). Although the logit and probit are sometimes viewed as interchangeable, they each make different assumptions about functional form—the probit

uses normality and the logit uses log normality. All of the features of the Heckman estimator are based on the assumption of bivariate normality and therefore require the use of the probit. Although previous research has used logit estimation, Bushway et al., (2007) point out that this is incorrect. The selection model is specified using probit regression and estimated using maximum likelihood (Greene, 1995, 2000) as follows:

$$s_i^* = \alpha x_i + \phi z_i + \mu_i$$

$$s_i = 1 \; if \; s_i^* > 0, s_i = 0, otherwise$$

$$y_i \; observed \; if \; s_i = 1$$

where s_i^* is an unobserved latent variable that determines the likelihood of survey participation for individual *i*, and s_i^* depends on a vector of observed characteristics x_i, a vector of exclusion restrictions z_i, and random error μ_i. If there is no correlation in the error terms, the expected outcome for individuals who do not participate in the online survey depends only on the observed characteristics x_i. The characteristics of the unobserved participants are the same as those that participated. However, if online participation in the Stress Test and depression diagnosis is correlated, the information that someone does not participate in the online survey changes the conditional distribution of ε_i for that person and the likelihood that he or she is diagnosed with depression. The error term, ε_i becomes biased.

The parameter ρ measures the correlation between the error terms of the substantive and selection models (Barnighausen et al., 2011; Clark and Houle, 2014). ρ has a potential range between -1 and +1 and gives some indication of the likely range of selection bias. A correlation with an absolute value of 1 would occur if the regression coefficients of the selection model and the regression coefficients of the substantive model were estimated by

identical processes (i.e., potential selection bias). Conversely, a value of ρ closer to zero would suggest that data are missing randomly or the regression coefficients of the selection model and the regression coefficients of the substantive model were estimated by unrelated processes (i.e., less evidence of selection bias) (Barnighausen et al., 2011; Clark and Houle, 2014). With regard to this paper, a significant negative ρ indicates that persons who do not participate in the Stress Test are more likely to be diagnosed with depression than Stress Test participants. A significant positive ρ Indicates that Stress Test respondents are less likely to be diagnosed with depression. The Wald test of independent equations assesses the null hypothesis that listening to BBC Radio 4 and the likelihood to be diagnosed with depression are independent of each other, i.e., whether ρ is significantly different from zero. Section 4 outlines the results of the model specified above.

4. Results

Table 3 presents the results of the Two Step Bivariate Probit model with depression the dependent variable in the outcome model and South of England residency as a proxy for listening to BBC Radio 4 in the selection model. This confirms the absence of selection bias in the Stress Test. The coefficient ρ was negative -0.22, indicating that persons who did respond to the Stress Test are less likely to be diagnosed with depression controlling for the IMD quintile of their LSOA. However, the Wald test indicates that this relationship is not significant (p=0.893). Examining each of the models in turn, a simple probit model was used to examine the outcome equation; the relationship between depression outcomes for individuals and area level deprivation. Using LSOA based IMD data, rather than the demographic and socio-economic variables in the Stress Test, ensures that there is data for both participants and non-participants. A binary probit model was used to estimate the effect of IMD quintile on depression. Furthermore, research in health geography and public health has found a significant relationship between area level deprivation and outcomes of

depression (Skapnakis et al., 2005; O'Campo et al., 2009; Gatrell and Elliott, 2009). Table 1 demonstrates that each IMD quintile is negatively associated with depression, relative to the most deprived LSOA. Thus, as with previous research, a simple probit model found that individuals diagnosed with depression by the Stress Test are significantly more likely to reside in deprived areas relative to the least deprived areas. A second simple probit model was used to estimate the effects of IMD quintile and rural residency on South of England residency, our proxy for BBC 4 Radio listenership. In the selection model each IMD quintile is linearly positively associated with living in the South of England, relative to the most deprived LSOA. There is a significant negative relationship between residing in the South of England and living in an area with a high deprivation score relative to the least deprived areas. This relationship is to be expected with the Greater London Area and South East regions the wealthiest across England (ONS, 2012).

Table 4 Two Step Bivariate Probit model for depression with depression as the dependent variable in the Outcome Equation and Southeast of England residency as the dependent variable in the Selection Equation

The "heckprob" command in STATA (StataCorp, 2013) was used to test for sample selection bias. The result of the Two Step model presented in Table 4 indicates that the relationship between IMD Quintile and diagnosis of depression remain the same. As with the simple probit model, moving from the most to the least deprived IMD quintile has a significant negative effect on whether an individual is case diagnosed with depression. However, the magnitude of the estimated coefficients for each quintile is lower in the simple probit model compared to the Two Step model. Individuals in the least deprived LSOA have a slightly lower probability of being diagnosed with depression relative to the most deprived LSOAs in the Two Step model compared to the simple probit model. This slight increase is intuitive. As

noted, the Greater London Area and the South of England regions, used as a proxy for BBC 4 Radio listenership are the wealthiest regions in England. Thus, as the descriptive statistics presented in Kinderman et al., (2013) indicate participants in the Stress Test are likely to have higher levels of education and income compared to non-participants. One may expect that decreases in deprivation (as one moves from Quintile 1 to Quintile 5) would have a smaller coefficient sign in the Stress Test sample than in a general population sample. That is, a disproportionately smaller sample of individuals in the most deprived LSOAs would mean that the size gradient for each increase in IMD Quintile may be underestimated relative to the most deprived IMD quintile. The largest difference is seen in Quintile 5, the least deprived group of LSOAs. Using the Two Step Bivariate Probit model, the IMD coefficient increases from -0.15 to -0.21 relative to the least deprived area. However, the estimated correlated coefficient, ρ indicates that sample selection bias is not a significant issue in the Stress Test data. Thus, although the Stress Test was launched on BBC Radio 4 and hosted online, the initial analyses of the sample produced estimates that appear to be generally accurate and free of selection bias. Thus, confidence in using the Stress Test for further population-based research is improved with the modelling of the bivariate probit sample selection model.

5. Discussion

Commenting on the explosion of Big Data over the last decade, Bell et al., (2009) argue that scientific research is now entering a 'fourth paradigm'. Bell et al., (2009) state that while earlier paradigms in social sciences were characterised by experimentation design and probability based sampling (Myers and Lorch, 2010), the latest approaches are strongly driven by the availability of data at an unprecedented scale. Data collected via bodies not necessarily interested in experimental design or data collected for reasons other than academic research may not be subject to the strict probability sampling required for

inferential statistics. This has to be taken into consideration when analysing data, calibrating models or testing hypotheses (Brunsdon and Comber, 2012; Birkin, 2013). Social science needs to find models that are able to robustly examine these new data sources. This is particularly true of web based surveys that appeal to the public interests; one has little control of who decides to respond to an online survey as the majority of web surveys allow participants to self-select into the sample (Schonlau, 2004). As many online surveys do not seek to meet externally defined demographic, socio-economic or spatial quotas based on national registers, the sample representativeness of these surveys needs to be tested before the data may be used for research. Using a Two Step Probit Selection Model to test the representativeness of the Stress Test, this paper did not find a statistically significant selection on unobserved factors for participants of the Stress Test. That is, the underlying characteristics of survey participants (those who found out about the survey, visited the site, completed registration and consented to the test) are not statistically different to non-participants on the substantive variables of interest, specifically area level deprivation. This study demonstrates the usefulness of web-based surveys for health research, however the representativeness of web surveys should be continuously monitored, particularly if policy relevant questions are being addressed.

Although sample selection methods are useful for detecting and statistically correcting selection bias, they do have limitations (Cuddleback et al., 2004). Any method for detecting and correcting selection bias is only as good as the selection model. Therefore, if the selection model is misspecified (e.g., important variables are missing from the model, only main effects are specified when interactions are present, or linearity is specified in the presence of non-linearity), methods of detecting and statistically correcting selection bias may be inaccurate or, unbeknownst to the researcher, may make estimates worse. Thus, Cuddleback et al., (2004) note that it can be helpful, though not always possible, for the

researcher to have some general idea about the source and direction of the bias *before* applying any methods of correcting selection bias, and some idea about the validity of the corrections *after* they are made. With regard to this paper, one may expect that decreases in deprivation (as one moves from Quintile 1 to Quintile 5) would have a smaller coefficient sign in the Stress Test sample than in a general population sample. That is, a disproportionately smaller sample of individuals in the most deprived LSOAs would mean that the size gradient for each increase in IMD Quintile may be underestimated relative to the most deprived IMD quintile. This pattern is observed in Table 1. The sign and significant level of each IMD quintile remains the same relative to the most deprived LSOAS. However, the magnitude of the coefficients for each quintile is lower in the simple probit model than in the Two Step Bivariate probit model. The largest difference is seen in Quintile 5, the least deprived group of LSOAs, where the IMD coefficient increases from -0.15 to -0.21 relative to the least deprived area. Thus, although these are only slight adjustments and sample selection bias is not a significant issue in the Stress Test, the changes in the coefficients are what one would expect given the profile of BBC Radio 4 listeners.

Given that most of the research in social work is subject to selection bias (Cuddleback et al., 2004; Bushway et al., 2007) and the increasing use of non-statutory data, researchers need to implement methods for detecting and correcting selection bias in their research. This paper explores the use of sample selection methods to detect and statistically correct for selection bias resulting from non-participation in a web-based survey. Recent research in public health and epidemiology (Chaix et al, 2011; Barnighausen et al, 2011) examining issues of sample selection bias each conclude that that *their* method of choice should be routinely used. However, similar to Geneletti et al., (2011), this paper contends that the specific method is not so important, although it should be appropriate, but that health

based research should follow the key principles of thinking about the selection process and assessing sensitivity to different assumptions.

References

Bärnighausen T, Bor J, Wandira-Kazibwe S, (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology; 22, 27–35.

BBC Radio Review (2010). BBC Radio Review, Value Partners, London.

Bell G, Hey T, Szalay A (2009) Beyond the Data Deluge. *Science* 323, 1297– 1298.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. American Sociological Review, 48 (3), 386-398.

Bernardo D, Curtis A (2013). Using Online and Paper Surveys: The Effectiveness of Mixed Mode Methodology for Populations Over 50, Research on Aging 35(2): 220-240.

Bethlehem J. (2010), Selection Bias in Web Surveys, International Statistical Review, 78, 2, 161–188, doi:10.1111/j.1751-5823.2010.00112.x

Blümel J, Chedraui P, Aedo,S, Fica J, Mezones-Holguín E, Barón G, Bencosme A, Benítez Z, Luz M. Bravo, Andrés Calle, Daniel Flores, María T. Espinoza, Gustavo Gómez, José A. Hernández-Bueno, Fiorella Laribezcoa, Mabel Martino, Selva Lima, Alvaro Monterrosa, Desiree Mostajo[a] Eliana Ojeda, William Onatra, Hugo Sánchez, Konstatinos Tserotas, María S. Vallejo, Witis S, Zúñiga M. (2015). Obesity and its relation to depressive symptoms and sedentary lifestyle in middle-aged women, Maturitas, 80(1), 100-105, doi:10.1016/j.maturitas.2014.10.007

Birkin M., (2013), Geoinformatics & Geostatistics: An Overview, *Geoinformatics & Geostatistics* 1(1). http://dx.doi.org/10.4172/2327-4581.1000e101.

Bowling A (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27, 281–91.

Brunsdon, C., Comber, L., 2012. Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach. *Geoinformatica* 16, 675–690.

Bushway, S., B. D. Johnson, and L. A. Slocum. 2007. Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology* 23, 151–78.

Byung-Joo Lee and Lawrence C. Marsh (2000). Sample selection bias correction for missing response observations, *Oxford bulletin of economics and statistics*, 62, 2.

Clark SJ, Houle B (2014) Validation, Replication, and Sensitivity Testing of Heckman-Type Selection Models to Adjust Estimates of HIV Prevalence. *PLoS ONE* 9(11): e112563. doi:10.1371/journal.pone.0112563.

Chaix B, Billaudeau, N., Thomas, F, Havard, S., Evans, D., Kestens, Y. and Bean, K., 2011. Neighborhood effects on health: correcting bias from neighborhood effects on participation. *Epidemiology*, *22*(1), pp.18-26.

Christensen H, Jorm AF, Mackinnon AJ (1999). Age differences in depression and anxiety symptoms: a structural equation modelling analysis of data from a general population sample. *Psychological Medicine* 29, 325–339.

Cuddeback, G., Wilson , E., Orme, J. G., & Combs-Orme, T. (2004). Detecting and correcting sample selection bias. *Journal of Social Service Research, 30*(3), 19–33.

Fleming, C and Bowden M. (2009). Web-based surveys as an alternative to traditional mail methods. *Journal of Environmental Management* 90, 284-292.

Leach L, Mackinnon A, Poyser C, Fairweather-Schmidt, A.K. (2015). Depression and anxiety in expectant and new fathers: longitudinal findings in Australian men, *The British Journal of Psychiatry* Jun 2015, 206 (6) 471-478; DOI: 10.1192/bjp.bp.114.148775

Gatrell, A.C., Elliott, S.E., 2009. Geographies of Health, second edition Wiley.

Goldberg, D., Bridges, K., Duncan-Jones, P., & Grayson, D. (1988). Detecting anxiety and depression in general medical settings. *British Medical Journal* 297: 897–899.

Gosling, SD. Vazire, S, Srivastava, S; John, O. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2), 93-104. http://dx.doi.org/10.1037/0003-066X.59.2.93

Greene, W. H. (1981). Sample selection bias as a specification error: Comment. *Econometrica*, 49 (3), 795-798

Hranov, L. (2007) Comorbid anxiety and depression: illumination of a controversy, International Journal of Psychiatry in Clinical Practice, 11:3, 171-189, DOI: 10.1080/13651500601127180.

Health Survey of England, (2011). Health Social Care and Lifestyle, Office of National Statistics, England.

Heckman J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and A Simple Estimator for Such Models. *Annals of Economic and Social Measurement* 5: 475-492.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. Econometrica, 46 (6), 931-959.

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica, 47 (1), 153-161

Hogan DR, Salomon JA (2012). Spline-based modelling of trends in the force of HIV infection, with application to the UNAIDS Estimation and Projection Package. *Sexually Transmitted Infections*, 88 (Suppl 2), i52–i57.

Index for Multiple Deprivation (2010) Department for Communities and Local Government, London.

Jorm, A.F., Windsor, T.D., Dear, K.B.G., Anstey, K.J., Christensen, H, & Rodgers, B. Age group differences in psychological distress: the role of psychosocial risk factors that vary with age. Psychological Medicine 2005; 35: 1253–1263.

Kinderman P, Schwannauer M, Pontin E, Tai S, (2011). The development and validation of a general measure of well-being: the BBC well-being scale, *Quality Life Resource*, 20:1035–1042

Kinderman P, Schwannauer M, Pontin E, Tai S (2013). Psychological processes mediate the impact of familial risk, social circumstances and life events on mental health, *PLoS ONE*, 8 (10), art. no. e76564, doi: 10.1371/journal.pone.0076564.

Kinderman P, Schwannauer M, Pontin E, Tai S, Jarman I, Lisboa P. (2015). Causal and mediating factors for anxiety, depression and well-being, The British Journal of Psychiatry Jun 2015, 206 (6) 456-460; **DOI:** 10.1192/bjp.bp.114.147553.

Kish, L. (1965). Survey sampling. New York: John Wiley & Sons, Inc.

Koloski, N, Smith N, Pachana N Dobson A. (2008). Performance of the Goldberg Anxiety and Depression Scale in older women, Age Ageing, 37 (4): 464-467, doi: 10.1093/ageing/afn091

Lee, B., & Marsh, L. C. (2000). Sample selection bias correction for missing response observations. Oxford Bulletin of Economics and Statistics, 62 (2), 305-323.

Magnavita N. (2015). Work-related symptoms in indoor environments: a puzzling problem for the occupational physician, *International Archive Occupational Environmental Health*, 88:185–196, DOI 10.1007/s00420-014-0952-7.

Mallinson S, Popay J. (2007) Describing depression: ethnicity and the use of somatic imagery in accounts of mental distress, Sociology of Health & Illness, 29(6) 857–871. DOI: 10.1111/j.1467-9566.2007.01048.x

Montero-Marin J, Piva Demarzo MM, Pereira JP, Olea M, García-Campayo J (2014). Reassessment of the Psychometric Characteristics and Factor Structure of the 'Perceived Stress Questionnaire' (PSQ): Analysis in a Sample of Dental Students. PLoS ONE 9(1): e87071. DOI: 10.1371/journal.pone.0087071

Morrissey, K, Williamson, P and Espuny Pujol, F (2015). A Multinomial Model for Comorbidity in England of Long-standing Cardiovascular Disease, Diabetes, and Obesity. *Health & Social Care In The Community*. DOI: 10.1111/hsc.12251

McKenzie' K Murray, A, Booth T (2013). Do urban environments increase the risk of anxiety, depression and psychosis? An epidemiological study, *Journal of Affective Disorders***,** 150(3), 1019–1024. doi:10.1016/j.jad.2013.05.032

Myers JL, Well A, Lorch RF (2010). *Research design and statistical analysis*, 3rd edn. Routledge, New York

O'Campo, P., C., Salmon, J. Burke. 2009. Neighbourhoods and mental well-being: What are the pathways? *Health & Place*, 5(1): 56-68.

Office of National Statistics (2012). National Statistics Postcode Lookup (2011 Census) User Guide: 2012 Edition, Office of National Statistics: London.

Office of National Statistics (2012). *Source: Wealth and Assets Survey,* Office of National Statistics: London.

Paykel, E., Abbott, R., Jenkins, R., Brugha, T. and Meltzer, H. (2000), Urban-rural mental health differences in Great Britain: findings from the National Morbidity Survey, *Psychological Medicine*, 30, 269-80.

Pontin, E. (2012). Research on pathways to mental health: testing the mediating psychological processes model of mental disorder. D.Clin. Thesis. University of Liverpool: UK.

Rentfrow PJ, Jokela M, Lamb ME (2015) Regional Personality Differences in Great Britain. PLoS ONE 10(3): e0122245. doi:10.1371/journal. pone.0122245

Rubin DB (1976). Inference and missing data. *Biometrika*; 63, 581–92. 21.

Sánchez Fernández J, Muñoz Leiva F, Montoro Ríos FJ (2012). Improving retention rate and response quality in web-based surveys. *Computers and Human Behaviour*, 28, 507–514.

Sartoni, A. (2003). An Estimator for Some Binary-Outcome Selection Models without Exclusion Restrictions, *Political Analysis* 11(2), 111-138.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S. and Miles, A., 2013. A new model of social class? Findings from the BBC's Great

British Class Survey experiment. *Sociology*, *47*(2), pp.219-250. **DOI:**

10.1177/0038038513481128

Schonlau, M., Zapert, K., Simon, L.P, Sanstad, K.H., Marcus, S.M., Adams, J., Spranca, Berry, S.H. (2004). A Comparison between Responses from a Propensity-Weighted Web Survey and an Identical RDD Survey, *Social Science Computer Review*, 22 (1), 128-138.

Sforza E, Saint Martin M, Barth JC, Roche F. (2016). Mood disorders in healthy elderly with obstructive sleep apnea: a gender effect, *Sleep Medicine* 19 (2016) 57–62.

Skapinakis, P., G. Lewis, R. Araya. (2005). Mental health inequalities in Wales, UK: multi-level investigation of the effect of area deprivation. *British Journal of Psychiatry*, 186: 417-422.

Strabac, Z., & Aalberg, T. (2011). Measuring political knowledge in telephone and web surveys: A cross-national comparison. *Social Science Computer Review, 29* (2), 175-192.

Surtees, P. G., Dean, C., Ingham, J. G., Kreitman, N. B., Miller, P. M. & Sashidharan, S. P. (1983). Psychiatric disorder in women from an Edinburgh community: associations with demographic factors. *British Journal of Psychiatry* **142**, 238-246.

StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

Stolzenberg, R. M., & Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. American Sociological Review, 62 (3), 494-507.

van de Ven W, van Praag B, (1981) The demand for deductibles in private health insurance: a probit model with sample selection*, Journal of Econometrics* 17, 229-252.

Yeager, D.S., Krosnick, J.A., Chang, L., Javitz, H.S., Levendusky, M.S., Simpser, A., Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples *Public Opinion Quarterly*, 75 (4), 709-747.

Table 1 Advantages and disadvantages of online surveys

Advantages	Disadvantage
Ease of data gathering	Non-universal access and or use of, the Internet across regions and demographic and socioeconomic profiles.
Minimal costs	Sample bias: non-random exclusion of individuals from the sample frame
Automation in data input and handling, thus reducing human error	Non response bias
Flexibility of design: For example dropdown boxes can present respondents with a range of possible answers, pop-up windows can provide additional information, questions can be ordered randomly, skip patterns may be built	Absence of interviewer
No geographical, linguistic or temporal limitations	Inability to reach challenging groups
Visual and audio stimuli can be incorporated	
Prompts can alert respondents if they skip or incorrectly answer questions,	

Table 2 Descriptive Statistics for English Respondents to the Stress Test

Age	
Average Age	43
Ethnicity	
White	96%
Education	
No GCSE	1.8%
GCSE	8.2%
Post 16 Vocational Course	2.2%
A levels	13%
Undergraduate	47%
Post graduate	27%
Occupational Status	
Still at school/university	5.7%
FT employment	55%
PT Employment	14%
Self Employed	9.8%
Unemployed	5%
Retired	8.5%
Voluntary	1.8%
Income	
Up to £9,999 (less than £199 p/w)	9.2%
£10,000 - £19,999 (£200-389 p/w)	15.5%
£20,000 – £29,999 (£390-579 p/w)	18%
30,000 – £39,999 (£580-769 p/w)	16%
£40,000 – £49,999 (£770-969 p/w)	12%
£50,000 – £74,999 (£970-1,449 p/w)	16%
£75,000 plus (£1,450 plus p/w)	13%
Relationship Status	
In a relationship (not married/not living together)	8%
Cohabiting	16%
Married (first marriage): 38	38%
Civil partnership:	0.68%
Divorced	6.8%
Divorced and remarried	7.33%
Separated (but still legally married)	2.54%
Widowed	1.7%
Widowed and remarried	0.4%
Single and never married	19%

Table 3 Cut-offs for Goldberg Depression by Age and Gender

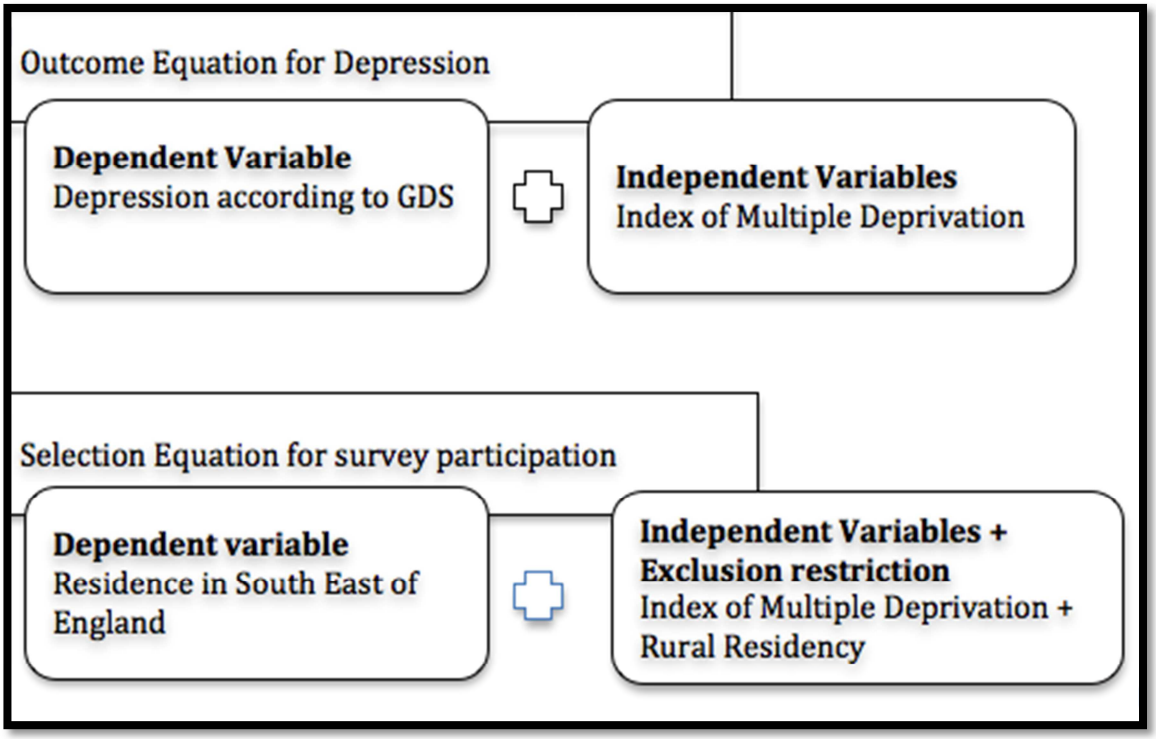
Age group in years	Males	Females
20–24	**6**	**6**
25–44	**6**	**5**
45–64	**5**	**4**

Table 4 Two Step Bivariate Probit model for depression with depression as the dependent variable in the Outcome Equation and Southeast of England residency as the dependent variable in the Selection Equation

VARIABLES	Probit Model	Probit Model	Heckman Two Step Model
Outcome Model			
2nd IMD Quintile	-0.10**		-0.13*
	(-0.162 - -0.031)		(-0.239 - -0.013)
3rd IMD Quintile	-0.12***		-0.17**
	(-0.188 - -0.061)		(-0.281 - -0.066)
4th IMD Quintile	-0.15***		-0.15**
	(-0.213 - -0.088)		(-0.255 - -0.043)
5th IMD Quintile	-0.15***		-0.21***
	(-0.213 - -0.088)		(-0.328 - -0.098)
Constant	-0.27***		0.05
	(-0.320 - -0.221)		(-0.423 - 0.516)
Selection Model			
2nd IMD Quintile		0.35***	0.25***
		(0.290 - 0.401)	(0.196 - 0.297)
3rd IMD Quintile		0.31***	0.22***
		(0.251 - 0.367)	(0.171 - 0.273)
4th IMD Quintile		0.29***	0.26***
		(0.228 - 0.346)	(0.208 - 0.309)
5th IMD Quintile		0.47***	0.37***
		(0.407 - 0.524)	(0.323 - 0.424)
Rural Residency		-0.49***	-0.38***
		(-0.547 - -0.429)	(-0.426 - -0.338)
Constant		-0.77***	-0.97***
		(-0.807 - -0.728)	(-1.008 - -0.934)
Rho	N/A	N/A	-0.22
			(-0.537 - 0.091)
Observations	18,720	20,593	Y – 18,720; Y – 20,593

CI in parentheses, *** p<0.001, ** p<0.01, * p<0.05

Figure 1 Heckman Style Sample Selection model and data from the Stress Test



Highlights

- The ‘Stress Test’ was developed by clinical psychologists and hosted on the BBC website
- The Stress Test created a large amount of mental health and population data
- As respondents self-selected to participate, survey may comprise a non-random sample.
- A model is presented to detect and statistically correct for selection bias.
- Web-based survey produced geographically representative sample of the English population.